基于主题模型的Web服务聚类与发现机制*

李　慧　胡云凤
(西安电子科技大学经济与管理学院　西安 710071)

摘要: 【目的】针对网络中海量的 Web 服务, 提出一种有效的 Web 服务聚类与发现方法。【方法】利用 BTM 学习整个 Web 服务描述文档集的隐含主题, 通过推理得出每个文档的主题分布, 并进行聚类。在此基础上, 创建一个快速的 Web 服务发现机制。【结果】与使用 LDA 和外部语料库等方法进行对比实验, 本文方法的查准率和标准折损累计增益均有所提高。【局限】仅考虑服务的功能信息, 未将服务的质量信息纳入算法。【结论】实验结果显示该方法可以更准确地发现符合用户需求的服务。

关键词: Web 服务　主题模型　聚类　发现
分类号: G350

1　引　言

目前正在步入面向服务的时代, SOA(Service Oriented Architecture)架构被广泛应用, 而 Web 服务逐渐成为实现 SOA 架构的主流技术。SOA 架构遵循发现、绑定、执行的服务模式, Web 服务由提供者发布在私有的或公共的互联网平台上, 用户在海量的 Web 服务中发现符合自己要求的 Web 服务, 与之进行绑定调用, 实现自己的目的。在这个过程中, 用户不需要了解服务的实现方式, 只需服务能够提供给用户满意的执行结果。互联网平台上发布的服务日益增多, 如何从海量的 Web 服务中发现用户满意的服务, 即从发布的 Web 服务描述中发现能满足用户期望的服务, 是实现面向服务架构关键的一环。

Web 服务描述文本篇幅较短、特征稀疏和信息量少, 根据词语的共现程度来度量相似性不可行。基于关键词发现 Web 服务, 完全依赖词语共现程度, 十分不准确。为了丰富 Web 服务描述文本, 一些语义 Web 方法被用于服务发现, 例如基于语义或者本体发现 Web 服务的方法[1-3]。但是, 建立和维护本体十分困难, 并且需要大量的人工干预[4]。此外, 在面对海量的 Web 服务时, 由于没有有效的分类机制, 很难快速有效地发现 Web 服务。

针对上述问题, 本文提出一种利用 BTM (Biterm Topic Model)[5]的 Web 聚类与发现方法。BTM 对整个语料库的词对生成过程建模, 从而学习整个语料库的主题分布和主题–词分布, 结合向量空间计算词的 TF-IDF 值, 可以推理得到每篇 Web 服务描述的主题分布, 进而对其聚类。Web 服务发现过程为: 获取请求服务的类别; 对该类别下的服务进行基于主题相似度的过滤, 大大缩小检索范围; 计算请求服务与 Web 服务之间的词向量相似度, 结合主题相似度和词向量相似度, 找到满足用户需求的服务集合。

2　相关工作

对 Web 服务发现的研究, 大量的工作投入在利用本体、词典发现的方法[1-2,6-8]中。文献[2]运用领域本体提出一种 Web 服务发现方法, 该方法通过本体中的概念距离计算服务请求和发布的服务之间的语义相似度。文献[9]对 Web 服务进行语义标注, 帮助发现 Web 服务。但是这类方法需要大量的人工干预, 依赖于本体的好坏及维护工作, 词库在某些领域的词汇量不足

通讯作者: 胡云凤, ORCID: 0000-0002-7342-3755, E-mail: 1540520650@qq.com。
*本文系中央高校基本科研业务费专项资金资助项目“大数据环境下基于主题模型的信息服务研究”(项目编号: JB160606)的研究成果之一。

和更新较慢也可能导致发现结果不准确。并且，此类方法前提要求服务发布方或请求方要提供相关的领域本体，而通常情况下，服务请求方是非专业用户，不能提供专业的本体，因此，该类方法的效率和通用性受到限制。此外，上述 Web 服务发现方法，由于没有有效的分类机制，在面对海量的 Web 服务时，不能实现实时匹配。

聚类是一个有效处理大量数据的方法，根据某一相似性标准重新组织数据，将数据分为不同的簇，能够实现快速的信息检索。Abramowicz 等[9]提出一种 Web 服务过滤和聚簇的方法，但是过滤机制是基于 OWL-S (Web Ontology Language for Service)描述的 Web 服务。依然存在依赖本体的缺陷。可获得的 Web 服务，大多都用 WSDL(Web Service Describe Language)描述，也有很多利用 WSDL 对 Web 服务分类。Nayak 等[10]将 Web 服务描述转化到多维词向量空间，利用两个向量之间夹角的余弦，计算两个服务之间的距离，对服务进行聚簇。这是一种基于数理统计的分类方法，可归一化处理大规模的文本集，但忽略了描述文本词项之间的语义关系，并且消耗的运行时间和存储空间随着文本集规模的增加而增加。Cassar 等[11]研究利用 PLSA(Probability Latent Semantic Analysis)和 LDA (Latent Dirichlet Allocation)挖掘 Web 服务描述的主题用于聚类，实验结果显示，LDA 模型在大规模服务集中，帮助自动服务发现的效果较好。LDA 由 PLSA 发展而来，Blei 等[12]引入 Dirichlet 先验分布扩展 PLSA 模型，提出 LDA 模型，通过发现隐含主题，可以处理大量的文本数据，进而对其进行分类。Aznag 等[13]在对 Web 服务描述文档进行预处理(特征提取、分词、去除停用词和词干还原等)后，利用 CTM (Correlated Topic Model)学习 Web 服务和服务请求的隐含主题，对其进行分类，匹配提供的服务与请求服务的相似性，得到最终候选 Web 服务集。该方法存在以下不足：Web 服务描述文本通常较短，类似于短文本，但该方法并没有对服务文本进行扩充，缺少足够的词频共现；仅用 Web 服务库作为训练集，规模较小，难以获得高质量的主题模型，以上两点导致很难学习出 Web 服务的真实隐含主题。另外，仅仅将 Web 服务归入其主题分布最大的那一类主题，分类不够精确。Aznag 等[13]使用的主题模型 CTM，由 Blei 等[14]提出，该模型引入对数正态分布取代 LDA 中的狄利克雷分布，CTM 模型先验参数中包含一个协方差矩阵，描述每对主题之间的相关性，协方差矩阵中参数的数量与主题数量的平方成正比。魏强等[15]使用 Word2vec 和 Relatedness 文本扩充方法，从 Wikipedia 中提取特征扩充 Web 服务描述文本，并以英文 Wikipedia 作为训练集，利用 HDP 非参数主题模型进行主题建模，提出 Signature 方法进行服务匹配，在一定程度上改进了服务发现效果，但是其服务匹配阶段，计算输入输出相似度时，采用精确匹配、嵌入匹配、包含匹配、交叉匹配和失败匹配 5 种类型，对概念相似度的区分太过简单，不适用于服务数量庞大的情况。

利用外部知识库，对短文本进行特征扩充是一种较常见的方法，但恰当适合的外部资料库不容易找到。而 Web 服务涉及很多方面，很难找到恰当适合的外部资料库，文献[15]使用 Wikipedia 作为外部资料库和主题模型训练集，Wikipedia 作为一种综合性的文本库，用做特征补充和主题模型训练，并不准确。Web 服务并不包含所有行业，在某些领域使用较为频繁，Web 服务数量也较多；但在另一些领域，则没有或很少有 Web 服务的使用。如果 Web 服务不存在或数量极少，综合性的外部资料库包含太多的关联信息，反而会使精确度下降。文献[16]发现在短文本分类时，使用一个外部知识库会使建模精确度下降。

本文提出使用 BTM 对 Web 服务描述文本进行建模，BTM 对整个语料库无序词对的生成过程建模，利用整个库的全部词频共现信息学习出隐含主题，解决了短文本因稀疏性而导致学习出的主题不准确的缺点。由于 BTM 学习出的是整个 Web 服务库的主题分布和主题–词分布，提出一种推理方法计算得出每个 Web 服务描述文档的主题分布。

3 基于 BTM 的 Web 服务聚类与发现

基于主题模型的 Web 服务聚类与发现框架如图 1 所示。Web 服务库的预处理结果作为 BTM 建模的输入，建模的输出为整个库的主题分布和主题–词分布，结合预处理得到的 VSM(Vector Space Model)，推理计算出每个文档的主题分布作为聚类的输入。当有服务请求时，预处理得到请求 r 的词向量，此时，前面计算得到的整个库的主题分布和词分布不变，不需重新建

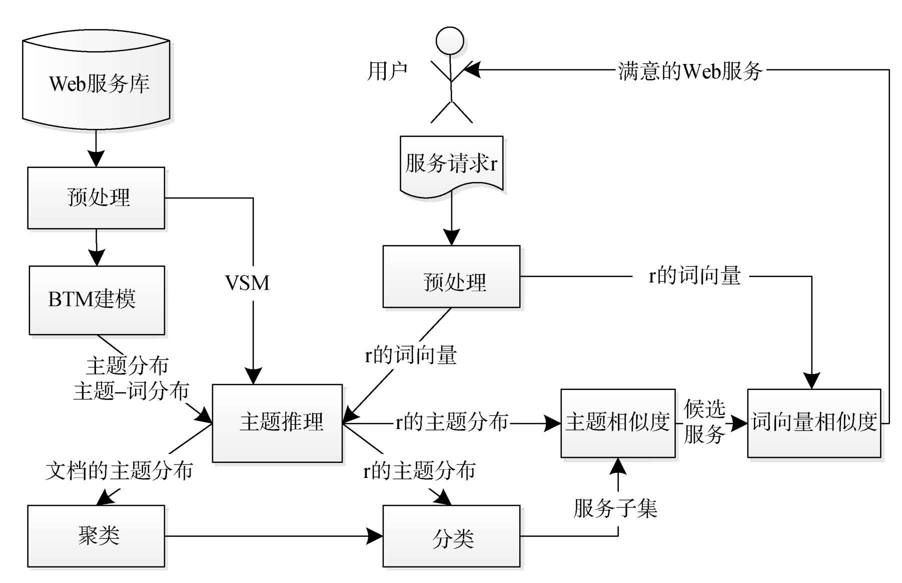


图 1　服务发现框架

模计算，直接利用其推理计算出请求 r 的主题分布并进行分类。得到 r 的类别后，计算主题相似度和词向量相似度得到用户满意的 Web 服务。

3.1　预处理

每一个 Web 服务都有一个使用 WSDL 描述的文档, WSDL 是一套基于 XML 的 Web 服务描述框架，经 W3C 对其进行标准化，有 1.1 和 2.0 两个版本，2.0 版本更为简单和实用。目前，1.1 版本使用较为广泛，但未来 2.0 版本可能会逐渐取代 1.1 版本。对使用 WSDL (1.1 版本和 2.0 版本)描述的服务文本数据预处理主要包含以下步骤:

(1) 特征抽取。从 WSDL 中抽取描述 Web 服务的全部特征，包括服务名称、服务功能文本描述、操作名称、输入/输出参数名称和参数类型等。

(2) 分词。提取出的文本中，有一些术语由几个单词组成，称为复合词，需要进行拆分。例如“BusinessDataCatalog”，将其拆分为“Business”，“Data”和“Catalog”。

(3) 去除标签和停用词。去除 Web 服务描述文本中的标签和停用词，以免在建模过程中影响准确度，利用 Standford POS Tagger 去除全部的标签和停用词，只有词性为名词、动词或形容词的会被留下来，去除的停用词，例如“a”，“is”和“that”等，去除的标签，例如“soap”，“type”和“binding”等 WSDL 标签。

(4) 词干还原。拥有相同词干的单词通常具有相同的意思，例如“recommended”只是“recommend”的过去式，利用 Porter Stemmer 对单词进行词干还原，使用词源形式的单词向量表示 Web 服务，更能有效地发现相关性。

(5) 词对抽取。BTM 与 LDA 直接建模在文档的词共现频率上不同，BTM 是基于整个语料库的词对共现率建模。初始短文本“BusinessData search for users”，在经过抽取、分词、去除标签停用词和词干还原后，抽取出的词对{(business, data), (business, search), (business, user), (data, search), …}，将整个库的词对作为 BTM 模型训练的输入。

(6)服务矩阵。抽取全部有用词后，计算词的 TF-IDF 值，利用 VSM(Vector Space Model)表示全部 Web 服务成为一个向量空间，将每一个 Web 服务表示为一个词向量，$s_i = \{w_{i1}, w_{i2}, \cdots, w_{in}\}$。其中每一个词的权重值由 TF-IDF 算出，$w_{ij} = tf_{ij} \times idf_j$，其中 tf_{ij} 为文档 i 中词 j 出现的频率, idf_j 为逆向文本频率，总文档数除以包含词语 j 的文档数的对数得到，一个词在文档中出现的频率越高，在其他文档中出现的频率低，则这个词有较高的重要性，权重值较大。

3.2　BTM 模型

BTM 通过统计整个语料库的词共现来建模学习隐含主题，不同于 LDA 是对单个文档中的词的生成过程建模，单个短文本缺乏足够的词频共现，LDA 建模结果并不稳定，BTM 对整个语料库的词对的生成

过程建模，整个语料库的词对的频率更稳定，也更能揭示出词之间的关系，学习出整个库的隐含主题。

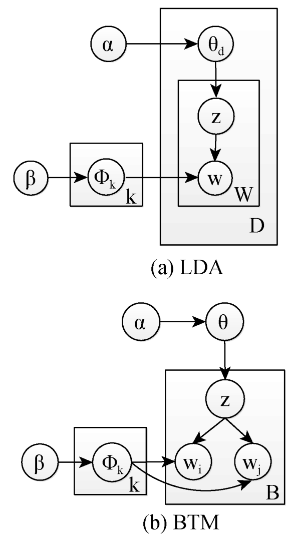


图 2　概率模型

LDA 建模一个文档的生成过程如图 2(a)所示：针对每一个文档随机生成一个主题分布θ_d，从θ_d中取样生成第 i 个词的主题 z，从 z 中再一步步采样生成词 w。可以看出，每一篇拥有一个主题分布，词的主题估计取决于同篇文档中其他单词。一元混合模型文档中所有的单词共享同一个主题 z，而 z 从全局主题分布θ中产生，假设整个语料库看作是主题的混合，从整个语料库统计信息，避免了短文本信息稀疏的问题，但是，一元混合模型假设一个文档仅有一个主题，并不符合实际情况，导致不能学习出好的主题。BTM 可以看做是一元模型和 LDA 的结合，如图 2(b)所示，BTM 假设一个全局的主题分布θ，但是其将每篇文档分割成词对，每对词对属于一个主题，BTM 允许一个文档有多个主题，即避免了一元混合模型的限制，又解决了 LDA 无法在短文本建模取得良好效果的问题。

3.3　Web 服务聚类

BTM 建模得出两个主要的参数，θ为 Web 服务集的主题分布，是一个 K 维的向量，K 为整个服务集的主题数目；Φ是主题–词分布矩阵，K 行 N 列，每一行是一个主题 Z 下的不同词的生成概率。由于 BTM 并未对每个文档的生成过程建模，所以不能直接得到每个文档的主题分布。可以通过θ和Φ推理计算得到，提出计算公式如下：

$$P(z_j \mid d)=\sum\nolimits_N P(z_j \mid w_i)\times P(w_i \mid d) \qquad (1)$$

其中，$P(w_i \mid d)$为 3.1 节中第(6)步计算得出的 TF-IDF 值。$P(z_j \mid w_i)$可以应用贝叶斯公式计算得出，计算公式如下：

$$P(z_j \mid w_i)=\frac{P(z_j)P(w_i \mid z_j)}{\sum\nolimits_K P(z_j)P(w_i \mid z_j)} \qquad (2)$$

其中，$P(z_j)$为主题分布θ中的主题 j 的概率，$P(w_i \mid z_j)$为主题–词分布Φ中，主题 j 下第 i 个词的概率。

得到文档的主题分布后，可以将不含语义信息的文档词向量表示形式转化为包含语义信息的主题向量表示，$s_i=\{P(z_1 \mid d_i),P(z_2 \mid d_i),\cdots,P(z_n \mid d_i)\}$。文本的主题向量表示形式，向量的每一项都是主题的概率，因此，文本相似度可以用 KL 散度(Kullback-Leibler Divergence)[17]计算，计算公式如下：

$$D_{KL}(p,q)=\sum_i p_i \ln\frac{p_i}{q_i} \qquad (3)$$

但是由于 KL 距离具有不对称性，即$D_{KL}(s_i,s_j)\neq D_{KL}(s_j,s_i)$。因此，使用 KL 距离的改进对称版本——JS 距离(Jensen–Shannon Divergence)[18]，其计算公式如下：

$$Sim_T(s_i,s_j)=D_{JS}(s_i,s_j)=\frac{1}{2}\left[D_{KL}\left(s_i,\frac{s_i+s_j}{2}\right)+D_{KL}\left(s_j,\frac{s_i+s_j}{2}\right)\right] \qquad (4)$$

将计算得出的 Web 服务描述文档相似性作为聚类算法的输入。

3.4　Web 服务发现

服务匹配是指在大量的服务中，能快速准确地查找到符合用户需求的候选服务。当输入一个服务请求时，对其进行数据预处理，通过主题分布和主题–词分布矩阵计算得出请求服务的主题分布，计算请求服务与各个聚簇中心服务的 JS 距离，将其归入距离最近的一类，即锁定了与服务请求类别相同的服务子集。

由于服务数量巨大，为了节约匹配时间，在得到类别相同的服务子集后，对子集中的 Web 服务基于主题进行过滤，利用 JS 距离计算服务请求的主题与子集中服务主题的相似度 $Sim_T(r,s)$，设定一个阈值，当相似度大于阈值时，将此服务加入候选服务集。计算候选服务集中的 Web 服务与服务请求的词向量相似度。

Web 服务词向量表示在 3.1 节中叙述, Web 服务与请求的词向量的相似度计算使用余弦距离, 余弦距离是向量空间相似度计算最常用的一种计算方法。计算向量空间中两个向量的夹角, 夹角越小, 则相似度越大。余弦距离计算公式如下:

$$\mathrm{Sim}_W(r,s)=\frac{\sum_{i=1}^{n} w_{ri}\times w_{si}}{\sqrt{\sum_{i=1}^{n}(w_{ri})^2}\sqrt{\sum_{i=1}^{n}(w_{si})^2}} \tag{5}$$

计算主题相似度, 可以得出服务间语义维度上的相似度, 而词向量相似度一定程度上反映了服务间统计层面的相似度。因此, 将计算得出的主题相似度和词向量相似度结合得出总的相似度, 即最终得出 Web 服务与请求服务的相似度如下所示:

$$\mathrm{Sim}(r,s)=\alpha \mathrm{Sim}_W(r,s)+(1-\alpha)\mathrm{Sim}_T(r,s) \tag{6}$$

其中, α为 Web 服务词向量相似度的权重, $0<\alpha<1$。

4 实验结果及分析

为对本文提出的方法进行验证, 本实验采用 WS-Dream[19]提供的数据集, 该数据集包含来自 69 个国家的 3 378 个 WSDL 文件, 15 811 个操作。本文利用 Weka 中的 KNN 算法[20]对 Web 服务进行聚类。使用查准率和标准折损累计增益进行效果评估。

查准率(Precision)是一种衡量检索出的全部结果中, 有用的结果比率有多大, 即检索出的相关的 Web 服务数量与检索出的全部 Web 服务数量之比, 公式如下:

$$\mathrm{Precision}=\frac{\text{相关Web服务}\cap\text{检索出的全部Web服务}}{\text{检索出的全部Web服务}} \tag{7}$$

查准率不考虑发现结果的位置信息, 仅能说明发现结果总体的质量高低。而折损累计增益(Discounted Cumulative Gain, DCG)统计方法对检索返回的每一个结果进行相关性等级排序, 相关性高的结果排序越靠前越好; 高相关性结果要比低相关性结果的贡献大很多。其公式如下:

$$\mathrm{DCG}_n=\sum_{i=1}^{n}\frac{2^{rel_i}-1}{\log_2(1+i)} \tag{8}$$

其中, rel_i是发现结果中排在第 i 位的结果的相关性等级。

不同发现方法发现的结果内容和数量不同, 为了能对不同发现方法进行对比, 可以使用标准折损累计增益(Normalize Discounted Cumulative Gain, NDCG)。其公式如下:

$$\mathrm{NDCG}_n=\frac{\mathrm{DCG}_n}{\mathrm{IDCG}_n} \tag{9}$$

其中, IDCG_n 是发现结果最优排序时, 计算出的 DCG_n。

实验中对三种主题学习方法用于服务发现的效果进行对比, 一种是 BTM, 一种是 LDA, 另一种是以 Wikipedia 作为外部知识库, 用 LDA 挖掘主题的方法。在 Java 环境下, 利用 JDK、Eclipse 和 JGibbLDA 等工具, 针对本文数据集的规模, 设置 20-100 个主题数, 不断调整主题数的大小, 进行迭代, 计算不同主题数对应的聚类结果的 F 值(F-measure), 得出在本文数据集中, BTM、LDA 和 LDA+Wiki 分别在 48、55 和 71 个主题数时达到聚类效果最优。表 1 列举 PAM 聚类中的一些主题, 以及相应主题中排序靠前的部分关键词。

表 1 部分主题–词分布

主题	词及其相应概率					
主题 1	cinema 0.048368	price 0.034546	version 0.030127	parameter 0.0187934	retrieve 0.009832	show 0.007656
主题 2	get 0.018766	service 0.01236	time 0.011016	result 0.009994	city 0.00875	hour 0.008847
主题 3	route 0.02820	location 0.02348	city 0.020753	weather 0.016753	tourist 0.014579	airplane 0.012782
主题 4	music 0.039748	album 0.018364	song 0.016545	release 0.018769	rock 0.013831	band 0.010342
主题 5	country 0.027491	british 0.018970	budget 0.017239	welfare 0.097230	culture 0.08371	party 0.052797

主题的词分布，能够很好地表现出每个主题的语义信息，主题 1、2、3、4 和 5 分别是相机、时间、旅游、音乐和国家信息，利用这些主题可以将多维的词向量降维到较低的主题向量，且具有语义信息和代表性，能够表示一个文档的特征。

在本次实验中，随机选取 12 个查询条件，每个查询条件的相关服务集由一组相关服务组成，每个相关服务有一个相关等级，相关性等级 $rel_i \in \{1,2,3\}$，3 表示高相关度，1 表示低相关度。发现结果与相关服务集进行对比，可以得到发现的相关服务数量和各个相关服务的相关性等级，从而计算出查准率和 NDCG。与经过预处理后，直接使用LDA 建模学习隐含主题并聚类发现 Web 服务，和经过基于 Wikipedia 扩充，再使用 LDA 建模的方法进行对比，结果如图 3 和图 4 所示。本文的方法用 BTM 表示，仅仅使用 LDA 建模的方法用 LDA 表示，经过扩充的方法用 LDA+Wiki 表示。

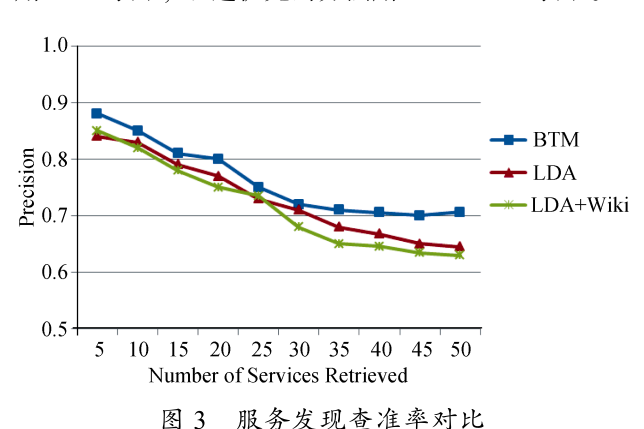


图 3　服务发现查准率对比

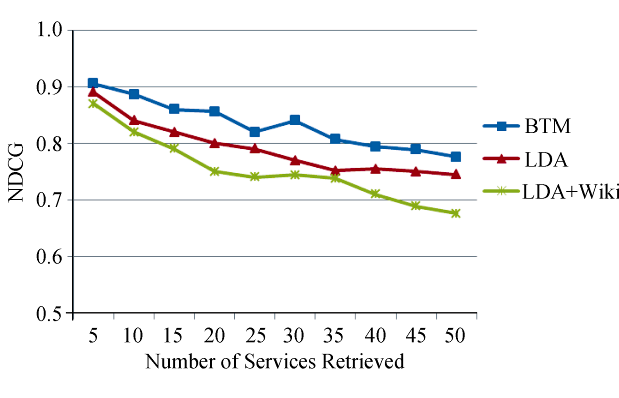


图 4　服务发现 NDCG 对比

关于查准率的对比如图 3 所示，LDA 和 LDA+Wiki 在检索服务数量分别大约为 10 和 25 的地方发生交叉，但总体上 LDA 要比 LDA+Wiki 更优一些。BTM 在整体上比另两种方法的查准率高一些，尤其当检索服务数量越多的时候，图中数量最大达到 50 时，BTM 与 LDA、LDA+Wiki 的差异达到最大，比 LDA 的查准率高 0.7%。NDCG 反映了发现相关结果的能力，如图 4 所示，BTM 优于 LDA 和 LDA+Wiki，大约比 LDA 高了 0.1%–0.8%，在服务数量最少为 5 时，高 0.1%，当服务数量达到 30 时，比 LDA 高 0.8%。以上说明 LDA 和 LDA+Wiki 由于查准率不高，错失了一些高相关性的 Web 服务。总体上，本文的发现方法更加符合 Web 服务描述文档的特点，能够帮助更好地发现 Web 服务，无论是查准率或标准折损累计增益，效能都更优。BTM 在整个 Web 服务描述库基础上全局建模，充分利用整个库的语义信息学习隐含主题，弥补了 Web 描述文本较短、缺乏词频共现和语义稀疏的特点。而 LDA 在文档层建模，很容易受到文档长度的影响，导致其学习出的主题不准确，不能够很好地表达语义信息，Web 服务发现的效果也不如 BTM。针对文本长度较短，利用 Wikipedia 进行扩充，基于扩充后的文本词向量使用LDA 进行建模，其服务发现的效果甚至不如直接使用LDA 进行建模，一方面由于外部资料库没有全部包含 Web 服务描述文档集的隐含主题；另一方面综合性的外部资料库拥有太多关联关系，包含各行各业的专有名词和过于丰富的词语，导致发现的查准率下降。

5　结　语

可获得的 Web 服务数量与日俱增，为了在海量的 Web 服务中快速有效地发现符合用户需求的服务，需要将功能类似的 Web 服务进行聚类。而 Web 服务描述文档较短，去除标签和停用词后，所剩的特征词并不多，利用 LDA 学习隐含主题，进而进行聚类，由于 LDA 对文档的生成过程建模，严重依赖于文本的长度，所以利用 LDA 对 Web 服务的聚类效果并不理想，不能很好地帮助服务发现。针对此问题，利用 BTM 学习 Web 服务文档集的隐含主题，推理出每个文档的主题分布，利用 JS 距离计算出各个文档之间的相似度，作为 KNN 算法的输入，对 Web 服务进行聚类。在 Web 服务发现阶段，综合主题相似度和词相似度发现 Web 服务。本文的发现方法充分利用整个 Web 服务库的语义资源，学习出较为准确的隐含主题，不需借助外部

知识库，减少了外部知识库因相关关系太多等原因带来的噪声信息。推理出各个描述文档的主题并进行聚类，对服务请求进行类别识别，大大缩小了查询范围，提高了查询效率。实验表明，本文提出的 Web 服务发现方法在准确性上具有一定的优越性。但是，该方法仅仅考虑 Web 服务的功能，没有将服务质量纳入计量范围, Web 服务真正的执行率不能保证。下一步的工作主要对服务质量进行计算，将服务价格、可靠性和响应时间等与服务发现结合，为用户提供更加可靠的服务。另外，如何对每个 Web 服务进行标签标注，使用户在选择服务时能够一目了然，选择符合要求的服务进行仔细研读，也是未来的一个方向。

参考文献：

[1] Farrag T A, Saleh A I, Ali H A. Semantic Web Services Matchmaking: Semantic Distance-based Approach [J]. Computer and Electrical Engineering, 2013, 39(2): 497-511.

[2] Lu G, Wang T, Zhang G, et al. Semantic Web Services Discovery Based on Domain Ontology [C]. In: Proceedings of the 2012 World Automation Congress (WAC). 2012: 1-4.

[3] 石敏，赵文栋，张磊．一种基于本体划分的语义 Web 服务发现算法[J]．计算机工程，2014，40(2): 175-179. (Shi Min, Zhao Wendong, Zhang Lei. A Semantic Web Service Discovery Algorithm Based on Ontology Partition [J]. Computer Engineering, 2014, 40(2): 175-179.)

[4] Atkinson C, Bostan P, Hummel O, et al. A Practical Approach to Web Service Discovery and Retrieval[C]. In: Proceedings of the 2007 IEEE International Conference on Web Service. 2007: 241-248.

[5] Yan X, Guo J, Lan Y, et al. A Biterm Topic Model for Short Texts [C]. In: Proceedings of the 22nd International World Wide Web Conferences. 2013: 1445-1456.

[6] Qu M, Liu S, Bao T. On the Trusted Ontology Model for Evaluating the Semantic Web Services[C]. In: Proceedings of the 14th International Conference on Computer Supported Cooperative Work in Design.2010: 368-369.

[7] Kopecký J, Vitvar T, Bournez C, et al. Semantic Annotations for WSDL and XML Schema [J]. IEEE Internet Computing, 2007, 11(6): 60-67.

[8] 杨惠荣，刘珊珊，尹宝才，等．基于语义距离的 Web 服务匹配算法[J]．北京工业大学学报，2011，37(4)：591-595. (Yang Huirong, Liu Shanshan, Yin Baocai, et al. Matching Algorithm of Services Based on Semantic Distance [J]. Journal of Beijing University of Technology, 2011, 37(4): 591-595.)

[9] Abramowicz W, Haniewicz K, Kaczmarek M, et al. Architecture for Web Services Filtering and Clustering [C]. In: Proceedings of the 2nd International Conference on Internet and Web Applications and Services.2007.

[10] Nayak R, Lee B. Web Service Discovery with Additional Semantics and Clustering [C]. In: Proceedings of the 2007 IEEE/WIC/ACM International Conference on Web Intelligence. 2007: 555-558.

[11] Cassar G, Barnaghi P, Moessner K. Probabilistic Methods for Service Clustering [J]. In: Proceeding of the 4th International Workshop on Service Matchmaking & Resource Retrieval. 2010.

[12] Blei D M, Ng A Y, Jordan M I. Latent DirichletAllocation[J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[13] Aznag M, Quafafou M, Rochd E M, et al. Probabilistic Topic Models for Web Services Clustering and Discovery[A]. // Service-Oriented and Cloud Computing[M]. Springer-Verlag Berlin Heidelberg, 2013.

[14] Blei D M, Lafferty J D. Correlated Topic Models[C]. In: Proceedings of the 23rd International Conference on Machine Learning. 2005.

[15] 魏强，金芝，许焱．基于概率主题模型的物联网服务发现[J]．软件学报，2014，25(8): 1640-1658. (Wei Qiang, Jin Zhi, Xu Yan. Service Discovery for Internet of Things Based on Probabilistic Topic Model [J]. Journal of Software, 2014, 25(8): 1640-1658.)

[16] Zhu Y, Li L, Luo L. Learning to Classify Short Text with Topic Model and External Knowledge[A]. //Knowledge Science, Engineering and Management[M]. Springer Berlin Heidelberg, 2013.

[17] Duda R O, Hart P E, Stork D G. 模式分类[M]．李宏东，姚天翔等译．第 2 版．机械工业出版社, 2003. (Duda R O, Hart P E, Stork DG. Pattern Classification [M]. Translated by Li Hongdong, Yao Tianxiang, et al. The 2nd Edition. China Machine Press, 2003.)

[18] Lin J. Divergence Measures Based on the Shannon Entropy [J]. IEEE Transactions on Information Theory, 1991, 37(1): 145-151.

[19] Zhang Y L, Zheng Z B, Lyu M R. A QoS-aware Search Engine for Web Services [C]. In: Proceedings of the 8th International Conference on Web Services. Miami, Florida, USA. 2010.

[20] Cover T M, Hart P E. Nearest Neighbor Pattern Classification[J]. IEEE Transactions on Information Theory, 1967, 13(1): 21-27.

作者贡献声明:

李慧: 提出研究思路和实验过程, 修改论文;
胡云凤: 提出具体方案, 完成实验, 撰写论文。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: 1540520650@qq.com。

[1] 李慧, 胡云凤. wsdldataset.rar. 由 WS-Dream 提供的 Web 服务数据.
[2] 李慧, 胡云凤. stopwords.dat. 在数据堂下载的原始停用词表基础上, 加入 WSDL 标签等停用词.
[3] 李慧, 胡云凤. pre-result.dat. 经预处理后的 Web 服务.
[4] 李慧, 胡云凤. topic.txt. 建模得到主题–词分布.
[5] 李慧, 胡云凤. simT.dat. Web 服务相似度矩阵.

收稿日期: 2015-12-22
收修改稿日期: 2016-02-03

Clustering and Discovering Web Services with Topic Model

Li Hui　Hu Yunfeng
(School of Economics and Management, Xidian University, Xi'an 710071, China)

Abstract: [**Objective**] We propose an effective method to cluster and discover the needed Web services. [**Methods**] First, we employed the Biterm Topic Model to learn the latent topics of the Web service description corpus. Second, we retrieved and clustered each document's topic distribution. Finally, we created a mechanism to discover Web service quickly. [**Results**] The proposed method achieved better precision rate and normalized discounted cumulative gain than methods using Latent Dirichlet Allocation and external corpus. [**Limitations**] Only considered functions of the Web services, and did not include the quality factors to the algorithm. [**Conclusions**] The proposed method could identify the needed services more accurately.

Keywords: Web service　Topic model　Clustering　Discovery